Intensive Care Med (2016) 42:3–15
DOI 10.1007/s00134-015-4124-3

SYSTEMATIC REVIEW

Florence C. M. Reith
Ruben Van den Brande
Anneliese Synnot
Russell Gruen
Andrew I. R. Maas

The reliability of the Glasgow Coma Scale: a systematic review

Received: 15 July 2015
Accepted: 26 October 2015
Published online: 12 November 2015

Take-home message: The overall reliability of the GCS is adequate, but can be improved by a renewed focus on adequate training and standardization. The methodological quality of reliability studies should be improved.

Electronic supplementary material
The online version of this article (doi:10.1007/s00134-015-4124-3) contains supplementary material, which is available to authorized users.

F. C. M. Reith (✉) · R. Van den Brande · A. I. R. Maas
Department of Neurosurgery, Antwerp University Hospital, Wilrijkstraat 10, 2650 Edegem, Belgium
e-mail: florence.reith@uza.be
Tel.: +3238214401

F. C. M. Reith · R. Van den Brande · A. I. R. Maas
University of Antwerp, Edegem, Belgium

A. Synnot
Australian & New Zealand Intensive Care Research Centre (ANZIC-RC), School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

A. Synnot
Cochrane Consumers and Communication Review Group, Centre for Health Communication and Participation, School of Psychology and Public Health, La Trobe University, Melbourne, Australia

R. Gruen
Central Clinical School, Monash University, Melbourne, Australia

R. Gruen
Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

A. Synnot
ANZIC-RC, Department of Epidemiology and Preventive Medicine, Monash University, The Alfred Hospital, Level 6, 99 Commercial Road, Melbourne, VIC 3004, Australia

R. Gruen
Central Clinical School, Level 6, The Alfred Centre, 99 Commercial Road, Melbourne, VIC 3004, Australia

Abstract *Introduction:* The Glasgow Coma Scale (GCS) provides a structured method for assessment of the level of consciousness. Its derived sum score is applied in research and adopted in intensive care unit scoring systems. Controversy exists on the reliability of the GCS. The aim of this systematic review was to summarize evidence on the reliability of the GCS. *Methods:* A literature search was undertaken in MEDLINE, EMBASE and CINAHL. Observational studies that assessed the reliability of the GCS, expressed by a statistical measure, were included. Methodological quality was evaluated with the consensus-based standards for the selection of health measurement instruments checklist and its influence on results considered. Reliability estimates were synthesized narratively. *Results:* We identified 52 relevant studies that showed significant heterogeneity in the type of reliability estimates used, patients studied, setting and characteristics of observers. Methodological quality was good ($n = 7$), fair ($n = 18$) or poor ($n = 27$). In good quality studies, kappa values were ≥0.6 in 85 %, and all intraclass correlation coefficients indicated excellent reliability. Poor quality studies showed lower reliability estimates. Reliability for the GCS components was higher than for the sum score. Factors that may influence reliability include education and training, the level of consciousness and type of stimuli used. *Conclusions:* Only 13 % of studies were of good quality and inconsistency in reported reliability estimates was found. Although the reliability was adequate in good quality studies, further improvement is desirable. From a methodological perspective, the quality of reliability studies needs to be improved. From a clinical perspective, a renewed focus on training/education and standardization of assessment is required.

Keywords Glasgow Coma Scale · Glasgow Coma Score · Grading scales · Reliability · Reproducibility of results · Systematic review

Introduction

The Glasgow Coma Scale (GCS), introduced in 1974, was the first grading scale to offer an objective assessment of the consciousness of patients [1]. The assessment of motor, verbal and eye responses of the GCS characterizes the level of consciousness. The picture provided by these responses enables comparison both between patients and of changes in patients over time that crucially guides management. The three components can be scored separately or combined in a sum score, ranging from 3 to 15. The sum score was initially used in research, but later also in clinical settings, even though summation of the three components incurs loss of information [2]. Both the GCS and the sum score are used in the intensive care unit (ICU) in a broad spectrum of patients with reduced level of consciousness and the sum score is integrated in several ICU classification systems [3–5]. An approximately linear relationship exists between decreasing sum scores and increasing mortality in patients with traumatic brain injury (TBI) [6], and the motor component is a strong predictor of poor outcome in moderate/severe TBI [7].

Reliable scoring is fundamental to the practical utility of the GCS. Conceptually, reliability is the degree to which an instrument is free from measurement error [8]. It has an external component (i.e. inter-rater reliability) which assesses the same subjects by different raters, and an internal component (i.e. intra-rater and test–retest reliability), which reflects the degree to which the scale yields identical results on different occasions and over time, assuming stable conditions [9]. Reliability is, however, not an inherent property of a test, but a characteristic of the scores obtained when applying the test [10]. Estimates of reliability are influenced by test properties, rater characteristics, study settings, heterogeneity of subjects and how subjects are treated, e.g. by intubation and sedation. It is important to identify factors that are potentially modifiable in order to improve the applicability of the GCS.

The reliability of the GCS has been examined in many studies, using a variety of measures, but remains an area of some controversy [11]. Various reports, specifically in the field of intensive care and emergency medicine, have criticized the GCS and questioned its general applicability [12–15]. Many assumptions are, however, based on limited evidence and mainly reflect personal opinions. No comprehensive systematic review on the reliability of the GCS and the factors that affect its reliability has been conducted since 1996 [16]. The aim of this systematic review is to explore the reliability of the GCS and the sum score, to identify influencing factors and to formulate recommendations for optimizing its reliability.

Methods

A protocol for this review was registered on PROSPERO (ID: CRD42014009488). The focus of the review was narrowed to reliability after publication of the protocol, but methods were followed as per protocol. We adhered to reporting and conduct guidance based on the preferred reporting items for systematic review and meta-analysis (PRISMA) [17] statement.

Eligibility criteria

Studies considered for inclusion were observational studies, such as cohort studies and case–control studies. We excluded case reports, letters, editorials or reviews. Studies were included if they used the GCS to verify the level of consciousness, and quantified its reliability by any statistical measure. We excluded studies in which GCS assessment was not obtained by physical examination of patients. Studies in which a majority (i.e. >50 %) of participants were assessed by the pediatric GCS were excluded.

Search strategy

A systematic literature search was executed from 1974 to January 2015 in MEDLINE, EMBASE and CINAHL. We developed search strategies using keywords and MeSH terms on the GCS and its clinimetric properties including reliability, validity, prognostic value and responsiveness (Table S1). In addition, the reference lists of eligible articles were screened for further relevant studies and relevant systematic reviews scanned for appropriate references.

Data selection and extraction

Citations were downloaded into Covidence (www.covidence.org), a software platform that manages the review process. An eligibility checklist was developed in accordance with the inclusion criteria. Two authors (F.R. and R.V.d.B.) independently reviewed all titles and abstracts. Potentially eligible articles were exported into the reference program Zotero (http://zotero.org). At this stage, the selected articles were screened again to identify articles relevant to the reliability of the GCS. Articles retained were obtained in full text and examined independently. Results were compared and disagreements were resolved by discussion. Data extraction was performed independently using a standard extraction form. The studies were subsequently screened for reporting factors that could influence the reliability of the GCS.

Assessment of methodological quality

The methodological quality of each study was assessed using the consensus-based standards for the selection of health measurement instruments (COSMIN) checklist [8]. This checklist evaluates studies on measurement properties of health measurement instruments. We used domains relevant to reliability (box A for internal consistency and box B for inter-rater reliability). The boxes contain standards on design requirements and statistical methods (Table S2). The assessment of a measurement property is classified as excellent, good, fair or poor based on the scores of the items in the corresponding box. An overall score is obtained by taking the lowest score for any of the items in the box. Assessment of quality was performed independently by two authors (F.R., R.V.d.B.) with disagreement resolved by discussion. The implications of methodological quality of studies on reliability estimates were considered by reporting results differentiated for quality ratings.

Data synthesis

Studies were grouped according to the statistical measures used. Within these groupings, the characteristics of each study and the methodological quality are described and presented in tabular form. Reliability measures are presented as reported by the authors and are differentiated, where possible, by the GCS components or the sum score. Where studies reported more than one reliability estimate (i.e. in different observer or patient populations), we included all estimates. Meta-analysis was explored but considered inappropriate due to high heterogeneity between studies.

The reported reliability measurements, including kappa, intraclass correlation coefficient (ICC), disagreement rate (DR) and Cronbach's alpha, have different properties and standards. The kappa statistic quantifies inter-rater reliability for ordinal and nominal measures. According to the classification system of Landis and Koch [18], kappa values between 0.00 and 0.20 indicate poor, 0.21 and 0.40 fair, 0.41 and 0.60 moderate, 0.61 and 0.80 substantial and >0.81 excellent agreement. A negative kappa represents disagreement. For reporting kappa values, we used cut-off values of 0.6 and 0.7, consistent with the recommendations by, respectively, Landis and Koch and Terwee et al. [19]. The ICC, expressing reliability for continuous measures, ranges from 0.0 to 1.0 with values >0.75 representing excellent reliability and values between 0.4 and 0.75 representing fair to good reliability [20]. The DR was developed at a time that kappa statistics were not in wide use in neuroscience [21]. It is expressed as the average distance from 'correct' rating divided by the maximum possible distance from correct rating. A lower DR reflects a higher reliability. Heron et al. [22] described that a low DR ranges from 0 to 0.299 whereas a high DR ranges from 0.3 to 0.5. Cronbach's alpha is the reliability statistic generally used to quantify internal consistency, which refers to the extent to which different items of a scale assess the same construct. Cronbach's alpha values >0.70 are considered adequate and values >0.80 as excellent. Alpha values >0.90 may indicate redundancy [23].

Results

After removing duplicates, 12,579 references were found in the literature search. After screening, 2896 citations were selected on the basis of their title/abstract for full text review. Of these, 71 were considered potentially eligible for this review. Twenty-four citations were excluded on full text (main reasons were inadequate study design, lack of data on the reliability of the GCS, irrelevance to the subject and pediatric population only). Cross-referencing and expert opinion identified eight further studies. The flow diagram (Fig. S1) summarizes this process. We included 52 studies, published in 55 reports.

Characteristics of studies

Of the 52 studies, published between 1977 and 2015, 6 were retrospective and 46 prospective (Table 1). The majority of studies were conducted in the ICU ($n = 22$) or emergency departments (ED) ($n = 12$), with the remainder in neurosurgical/neurological ($n = 9$), prehospital care ($n = 5$) and other ($n = 4$). Overall, 13,142 patients were assessed, with study sample sizes varying between 4 [24] and 3951 [25] patients. Three studies examined the GCS as part of the acute physiology and chronic health evaluation (APACHE) II score [26–28]. Standard errors or 95 % confidence intervals were rarely reported, limiting opportunities to synthesize estimates across different studies [10]. An extremely high level of heterogeneity ($I^2 > 98$ %) across studies reporting error estimates precluded a meaningful meta-analysis.

Methodological quality and reported estimates of reliability

The methodological quality of studies was evaluated as poor in 27 studies, fair in 18, and good in 7, while no study was rated as excellent (Table S3). Two studies were assessed using box A, as they measured reliability by means of internal consistency only [29, 30]. A total of six different statistical measures were identified to assess reliability (Table 2). Studies that did not report ICC,

Table 1 Characteristics of included reports ($n = 55$)

Study	Design[a]	Setting[b]	Patients[c]	Severity[d]	Treatment[e]	Observers[f]	GCS[g]	Reliability measure[h]	Quality of studies
Braakman et al. [47]	Pros	NeuroS	10 (video)	Sev	–	NS, R, N	M	Kappa	Poor
Teasdale et al. [21]	Pros	NICU	28 + 14 (video)	U	–	NS, Su, N	E, M, V, S	DR	Poor
Rimel et al. [48]	Pros	NeuroS	150, TBI	All 70 % Mi	–	R, N, EMT	Sr	% Of variation	Poor
Lindsay et al. [49]	Pros	–	126, SAH	GCS 6-15	–	NsR, Co	Sr	Kappa	Poor
Stanczak et al. [50]	Pros	NeuroS	101, GCS < 14	All	–	R, NPsy	E,M,V	Cα; P.r; Sp.r	Fair
Starmark and Heath [51]	Pros	ICU	26, Poisoned	Mo/Sev	I: 65 %	Authors	E, M, V, S	Unw.kappa	Poor
Starmark et al. [52]	Pros	NeuroS	47, NeuroP	All	I: 57 %	P, N	–	DisAG	Poor
Fielding and Rowley [35]	Pros	–	75, TBI/neuroP	U	–	N	E, M, V	% A, DR, ICC	Poor
Rowley and Fielding [34]	Pros	NeuroS	5/6, TBI/neuroP	U	–	N	E, M, V	% A; RC; DR	Poor
Tesseris et al. [53]	Pros	ICU	74	All	I: 18 %	P	S	Kappa	Fair
Ellis and Cavanagh [54]	Pros	NeuroS	12, (videotaped)	All	–	N, E	E, M, V	Cramér's V	Poor
Menegazzi et al. [24]	Pros	ED	4 (videotaped)	Sev	–	P, Pm	GCS	CC ($r_{1,2}$); kappa	Poor
Juarez and Lyons [33]	Pros	ICU	7 (videotaped)	U	I: U, S: 0 %	P, N	E, M, V, S	DR; kappa	Poor
Diringer and Edwards [29]	Pros	NICU	84, stroke/TBI	All	I: 48 %	Oct	E, M, V, S	Cα	Poor
Oshiro et al. [55]	Retr	NICU	291, aneurysm	All	–	US	Gr.sy	Kappa	Poor
Crossman et al. [56]	Pros	PH + ED	82	All	–	–	Scale	% Correct scores	Poor
Wijdicks et al. [57]	Pros	NICU	18	U	–	Ne, N, R	S	% A	Poor
Chen et al. [28]	Retr	ICU	342, db records	All	–	db, authors	A.p	% A, Kappa	Poor
Heron et al. [22]	Pros	ICU	6 (video)	All	I: 67 % S/P: 0 %	N	E, M, V, S	Kappa; DR	Poor
Lane et al. [58]	Pros	PH	4 (video)	GCS 5-15	I: 25 %	CA, PmS	E, M, V	RR improvement	Poor
Ely et al. [59]	Pros	ICU	38	U	I: 46 %	N, P, Npsy	S	W.kappa	Fair
Hollander et al. [25]	Pros	ED	3951, TBI	U	–	P, R	E4;M5;V5	Kappa; % A	Fair
Gill et al. [31]	Pros	ED	116	All	I: 16 %; P: 0 %	R	E, M, V, S	% A; kappa; Sp.r; Kτ	Good
Heard and Bebarta [40]	Pros	ED	39, poisoned	All 41 % Mi	P: 0 %	P	E, M, V, S	W.kappa	Fair
Wijdicks et al. [60]	Pros	ICU	120	All	I: 48 %;S/P: 0 %	N, R, P	E, M, V, S	W.kappa; Cα	Fair
Holdgate et al. [61]	Pros	ED	108, GCS < 15	All	S/I: 0 %	P; N	E, M, V, S	W.kappa; % A	Good
Baez et al. [62]	Pros	PH	4, Int.scen	GCS 6-15	I: U	P, EMS	C	dup.cas, sc.cor	Poor
Gill et al. [36]	Pros	ED	120	All	I: 18 %	P	E, M, V, S	% A; kappa; Sp.r; Kτ	Fair
Kerby et al. [63]	Retr	PH + ED	3052, BTP/PTP	All	I: 13 %	Pm, EDp	E, M, V,C	W.kappa	Poor
Kho et al. [26]	Pros	MS.ICU	37	U	–	Nr, Cr	E, M, V, S	ICC	Fair
Wolf et al. [64]	Pros	ICU	80	All	S: 0 %	N	E, M, V, S	Cα; ICC; w.kappa	Good
Nassar JR et al. [65]	Pros	MS.ICU	29	U	I: 35 %; S: 17 %	P, N, R, Ph	S	W.kappa	Poor
Akavipat et al. [66]	Pros	NeuroS	100, neuroP	All	I: 20 %; S/P: 0 %	P, N	E, M, V, S	Cα; ICC; w.kappa	Fair
Iyer et al. [67]	Pros	M.ICU	100	All	I: 45 %; S/P: 0 %	N, F, Co	E, M, V, S	ICC; Cα; w.kappa	Fair
Ryu et al. [68]	Retr	ED	876, TBI	Mi	–	E	Sr	Kappa	Poor
Stead et al. [69]	Pros	ED	69, neurolP	All	–	P, R, N	E, M, V, S	W.kappa; ICC	Good
Wenner et al. [27]	Retr	MS.ICU	95	U	–	db	A.p	Kappa	Fair
Fischer et al. [70]	Pros	M.ICU	267	All	I: 23 %; S: 20 %	Ne, N, P	E, M, V, S	% A; Cα; w.kappa	Good
Idrovo et al. [71]	Pros	SU	60, stroke	All	I: 3 %; S/P: 0 %	R, N	E, M, V, S	ICC; Cα; w.kappa	Good
Necioglu Orken [72]	Pros	NICU	124	U	I: 14 %;S/P: 0 %	Ne	S	Kappa	Fair
Ashkenazy et al. [73]	Pros	ICU	88	All	I: 100 %; P: 0 %	Nr	S	P.r	Poor
Bruno et al. [74]	Pros	ICU	176 TBI	Sev	I: 74 %;S/P: 0 %	N, NPsy, P, pR	E, M, V, S	W.kappa	Fair
Gujjar et al. [75]	*Pros*	*Wards*	*100*	*All*	*S: 0 %*	*N, R*	*Sr*	*Kappa*	–
Kevric et al. [76]	Pros	ED	203	All	I: U	P, N	E, M, V, S	% A; kappa	Fair
Namiki et al. [77]	Pros	ED	8 (video)	All	I: 0 %	R	Scale	% disAG	Poor
Patel et al. [78]	*Pros*	*NICU*	*51 TBI*	*All*	–	*Investigator*	*Overall*	*Cα*	–
Takahashi et al. [79]	Pros	ED	495, transported		–	P, N, R, Pm, St	S	W.kappa	Good
Winship et al. [80]	*Pros*	–	*4 (video)*	*All*	*I: 0 %*	*PmS*	*S*	*% A*	–

Table 1 continued

Study	Design[a]	Setting[b]	Patients[c]	Severity[d]	Treatment[e]	Observers[f]	GCS[g]	Reliability measure[h]	Quality of studies
Marcati et al. [81]	Pros	ICU/war	87, TBI	All	I:17 %;S/P:0 %	Ne, R	E, M, V, S	ICC;Cα; Sp.r; kappa	Fair
Sadaka et al. [30]	Pros	NICU	51, TBI	All	S: 0 %	Investigator	Overall	Cα	Poor
Winship et al. [82]	Pros	–	4 (video)	All	I: 0 %	PmS	E, M, V, S	% A	Poor
Benítez-Rosario et al. [83]	Pros	PC	156, AC	U	S: 19 %	P, R	S	W.kappa	Fair
Dinh et al. [84]	Retr	PH + ED	1181	U	I: 1 %; S: 28 %	Pm, EDp	S	ICC	Fair
Gujjar et al. [85]	Pros	Wards	100, medical	All	S/P: 0 %	Co, R, SHO	E, M, V, S	Kappa; Cα	Fair
Feldman et al. [32]	Pros	ED	9 (scenarios)	All	I: 0 %	EMT, Pm	E, M, V, S	% A	Poor

Italic: conference abstract or poster, excluded for analyses. Full reports of Gujjar et al. [75], Patel et al. [78] and Winship et al. [80] were published in Gujjar et al. [85], Sadaka et al. [30] and Winship et al. [82], respectively

[a] Design—*Pros* prospective cohort study, *Retr* retrospective cohort study

[b] Setting—*NeuroS* neurosurgical service, *ICU* intensive care unit, *NICU* neuro-ICU, *M.ICU* medical ICU, *MS.ICU* medical/surgical ICU, *ED* emergency department, *PH* pre-hospital; *SU* stroke unit, *PC* palliative care

[c] Patients—*TBI* traumatic brain injury, *SAH* subarachnoid hemorrhage, *neuroP* neurosurgical patients, *neurolP* neurological patients, *db* database; *Int.scen* Internet based scenarios, *BTP* blunt trauma patients, *PTP* penetrating trauma patients, *AC* advanced cancer patients

[d] Severity—All full range of severities of impaired consciousness. *Sev* severely impaired consciousness, *Mo* moderate impaired consciousness, *Mi* mild impaired consciousness, *U* unclear

[e] Treatment—*I* intubated patients, *S* sedated patients, *P* paralysed patients, *U* treatment was applied, but percentage of patients is unkown

[f] Observers—*NS* neurosurgeon(s), *R* resident(s), *N* nurse(s), *P* physician(s), *Su* surgeon(s), *NsR* neurosurgical registrar(s), *Co* consultant(s), *F* fellow(s), *Ne* neurologist(s), *EMT* emergency medical technician(s), *Pm* paramedic(s), *PmS* paramedical student(s), *Sn* nurse student(s), *St* medical student(s), *US* unspecified member of department of neurosurgery, *CA* conference attendees, *EDp* ED personnel, *Nr* research nurse(s), *Cr* research clerk(s), *Ph* physiotherapist, *NPsy* neuropsychologist(s), *SpR* specialist registrar(s), *E* expert viewer, *Oct* occupational therapy graduate students

[g] GCS—*E* eye, *M* motor, *V* verbal, *S* sum, *C* classification, *Sr* GCS sum range, *Gr.sy* GCS grading system, *A.p* GCS point in APACHE score

[h] Realibility measure—*DR* disagreement rate, *% A* percentage agreement, *DisAG* %disagreement, *W.kappa* weighted kappa, *Unw.kappa* unweighted kappa, *ICC* intraclass correlation coefficient, *Cα* Cronbach's alpha, *Sp.r* Spearman's rho, *Kτ* Kendall's τ-b, *CC* correlation coefficient, *RC* reliability coefficient, *dup.cas* correct identification of duplicated cases, *sc.cor* % of all scenarios scored correctly, *P.r* Pearson r

Table 2 Overview of reported reliability estimates differentiated by methodological quality of studies

	Good quality studies		Fair quality studies		Poor quality studies	
	No. of studies	No. of values	No. of studies	No. of values	No. of studies	No. of values
Kappa	7	81	15	143	10	41
ICC	3	9	5	23	1	5
Percentage agreement	3	14	3	16	8	89
Disagreement rate	0	0	0	0	5	123
Correlation coefficients	0	0	1	8	1	8
Cronbach's alpha	9	6	6	8	2	5

ICC intraclass correlation coefficient

kappa or Cronbach's alpha were rated at best as poor methodological quality, consistent with the COSMIN checklist. Similarly, use of unweighted kappa or having inadequate sample size precluded rating of highest quality (Table S2). Studies that were published soon after the introduction of the GCS were mostly judged to be of poor methodological quality.

Inter-rater reliability of the GCS

Kappa coefficient

A total of 265 individual kappa statistics were reported in 32 studies (Table 1). Often, it was not clarified whether weighted kappa statistics was applied. Methodological quality was good in 7 studies, fair in 15 and poor in 10. Figure 1 summarizes the reported kappa values in these studies, differentiated by quality rating. In the good ($n = 81$) and fair ($n = 143$) quality studies, 85 and 86 %, respectively, of all reported kappa values represented substantial reliability (Table S4). This percentage remained high at 78 and 67 % for kappa values $\geq$0.70. In the poor quality studies, 56 % of kappa's was $\geq$0.6. Of all reported 265 kappa values, 81 % showed substantial reliability (Table S4).

Considered both across and within the studies, there were no clear differences in kappa between the components (Table S5; Fig. 1). The sum score appeared generally less reliable than the components. Kappa values for the sum score represented substantial reliability in 77 % of reported estimates in good quality studies compared to 89, 94 and 88 % for the eye, motor and verbal components, respectively (Fig. 1). Kappa values reported in poor quality studies were lower. The studies that reported kappa for the GCS sum score as part of the APACHE II showed a mean of 0.34, representing fair reliability. Nevertheless, in the 16 studies performed in the ICU (Table S4), the sum score showed substantial agreement in 83 %, with higher scores for the components (90 % for eye score and 97 % for motor and verbal scores). Overall, in these ICU studies, 90.5 % of kappa's were $\geq$0.6.

Intraclass correlation coefficient

Nine studies reported ICC values (Table 3), of which eight were of good/fair quality. All ICC values (100 %) reported in the good quality studies ($n = 9$) were >0.75, representing excellent reliability. Kho et al. [26] reported ICC values for the GCS as part of the APACHE II score, with satisfying results except for the verbal component scores.

Percentage agreement

Fourteen studies expressed reliability by percentage agreement (Table S6). In the good ($n = 3$) quality studies, the percentage agreement ranges from 38 to 71 % for the sum score, and from 55 to 87 % for the components. Eleven studies were of poor/fair quality and confirmed lower percentages for the sum score. Some studies measured percentage agreement within the range of $\pm$1 point, which is considered more clinically relevant [31, 32]. In the absence of consensus on what level of percentage agreement is acceptable, the exact meaning of these numbers is unclear. One recent study, showing percentages ranging from 41 to 70 %, assessed this as low; however, 82 % of scores were within 1 point of correct scores [32].

Disagreement rate (DR)

The DR was used in five studies to express reliability of the GCS (Table S7). DR ranged between 0 and 0.143 and varied across the GCS components and sum score. The more recently published studies [22, 33–35] showed generally lower DR (i.e. higher reliabilities) than initially published by Teasdale et al. [21]. However, all studies were of poor methodological quality, limiting the strength of conclusions.

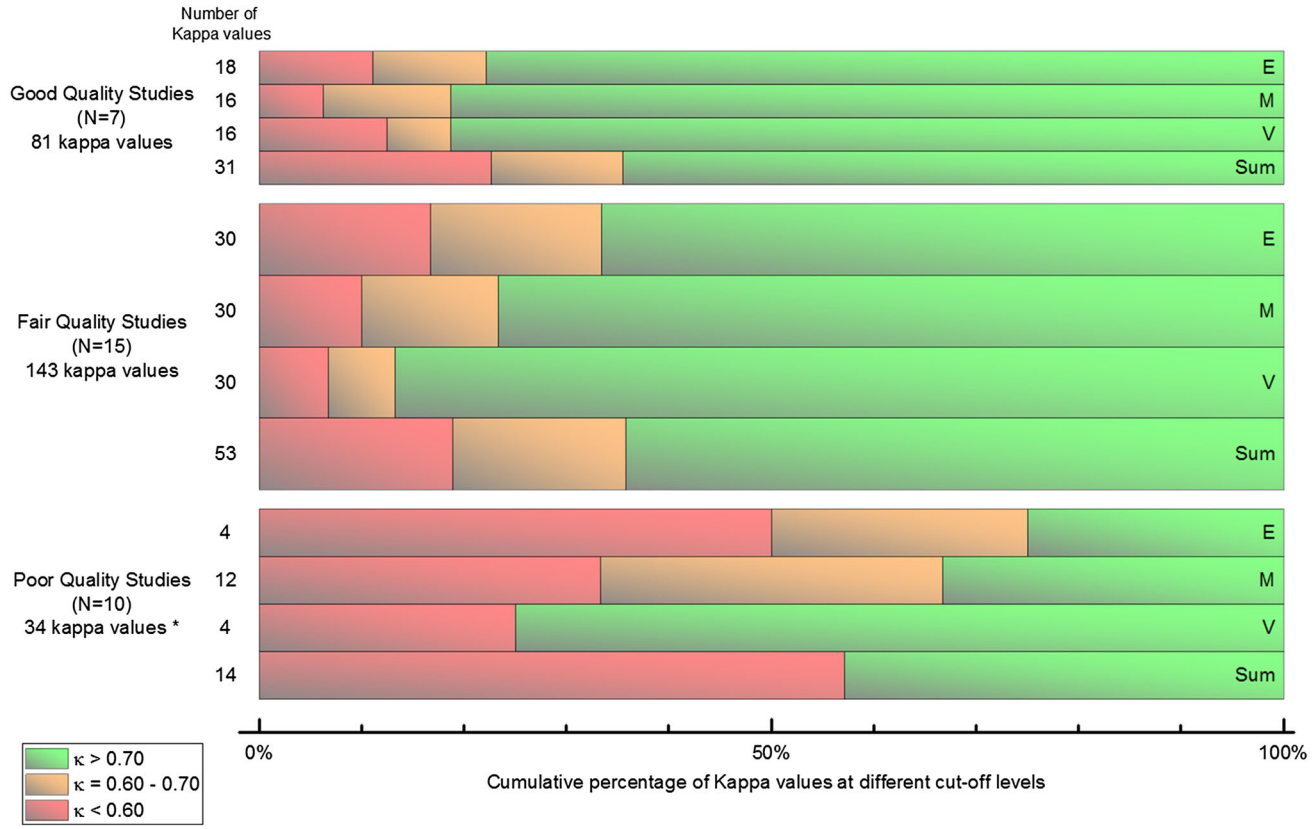

Fig. 1 Cumulative percentage of kappa values at different cut-off levels. *Asterisk* Seven kappa values reported in three poor quality studies were excluded from this figure as they represented reliability in a range of sum scores (categories)

Table 3 Intraclass correlation coefficients (95 % confidence intervals) for GCS components and sum score

Study	Pt[a] (N)	Setting	Eye	Motor	Verbal	Score
Good methodological quality						
Wolf et al. 2007 [64]	80	ICU	0.89	0.94	0.86	0.94
Stead et al. 2009 [69]	69	ED	0.934	0.921	0.911	0.964
Idrovo et al. 2010 [71]	60	SU	–	–	–	0.96 (0.93–0.97)
Fair methodological quality						
Kho et al. 2007 [26]	37[b]	ICU	0.83 (0.73–0.90)	0.87 (0.79–0.92)	0.40 (0.20–0.60)	0.83 (0.73–0.90)
			0.85	0.84	0.20	0.84
			0.86	0.86	0.40	0.80
			0.78	0.90	0.69	0.86
Akavipat et al. 2009 [66]	100	Ward	–	–	–	0.95[c]
Iyer et al. 2009 [67]	100	ICU	–	–	–	0.98 (0.98–0.99)
Marcati et al. 2012 [81]	87	ICU/W	–	–	–	0.99 (0.98–0.99)
Dinh et al. 2013 [84]	1181	ED				0.74 (0.37–1.12)
Poor methodological quality						
Fielding and Rowley 1990 [35]	75	–	L: 0.927 R: 0.895	L: 0.869 R: 0.855	0.974	–

ICU intensive care unit, *ED* emergency department, *SU* stroke unit, *ICU/W* ICU/ward, *L* left, *R* right

[a] Number of patients

[b] Observed by different rater pairs

[c] Mean of four values

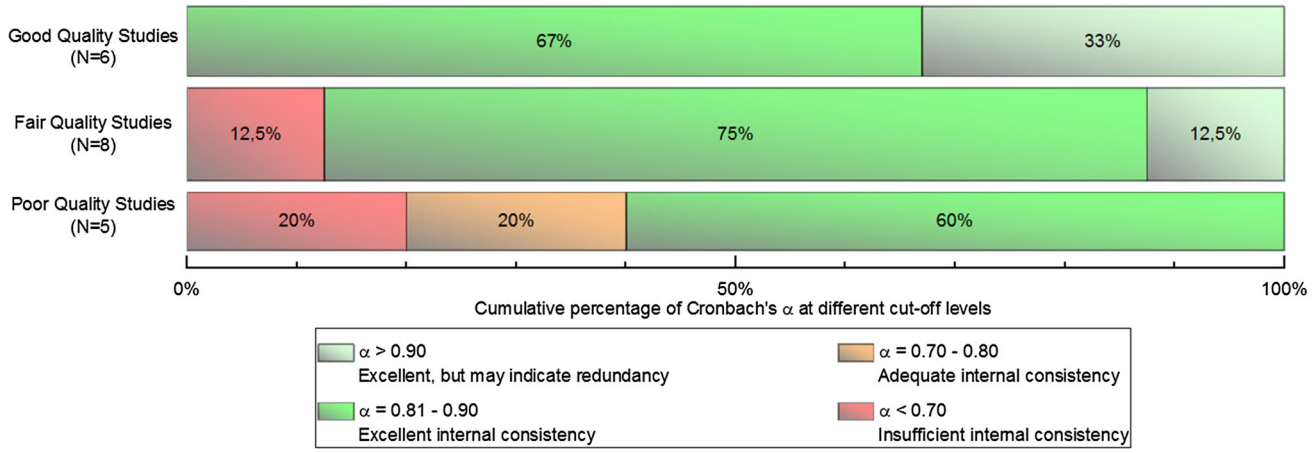

Fig. 2 Cumulative percentage of Cronbach's α at different cut-off levels

Correlation coefficients

Gill et al. [31, 36] reported correlation coefficients to assess pair-wise correlations between observations of two emergency physicians. The Spearman's rho ranged from 0.67 for the verbal score to 0.86 for the sum score. The Kendall rank ranged from 0.59 for the verbal score and 0.82 for the motor score. These measurements demonstrated moderate levels of agreement.

Intra-rater reliability and test–retest reliability

Five studies examined the intra-rater reliability of the GCS, but used different statistical tests and four were of poor methodological quality (Table S8). Clear conclusions can therefore not be drawn. Most authors of primary studies refrained from drawing clear conclusions. Only Menegazzi et al. stated that intra-rater reliability was high [24].

Internal consistency

Eighteen Cronbach's alpha values were reported in eleven studies (see Fig. 2; Table S9). Of the six values derived from good quality studies, 100 % are over 0.80, suggesting excellent internal consistency. Similar results are seen in the fair quality studies, but the poor quality studies show slightly less favorable results (60 % >0.80).

Overview of factors influencing the reliability of the GCS

Forty studies analyzed one or more factors that could influence the reliability of the GCS, identifying four observer-related factors described in 29 studies and three patient-related factors in 24 studies (Table 4). The beneficial role of training and education is supported by a majority of studies that assessed this influence, pointing to the potential for improvement of the reliability from training and education. The influence of the level of experience of observers appeared contradictory. The majority of studies investigating the influence of the type of profession showed similar reliabilities among different observer types. Some evidence suggests that the type of stimuli used to elicit a response in patients not responding spontaneously influences reliability. The consciousness level influenced reliability in the majority of studies, with higher agreements at the outer-ranges. We found conflicting results as to whether the type of pathology of patients influenced reliability and evidence on the influence of intubation and/or sedation on the reliability appeared inadequate. One primary study suggests that reliability of the verbal score is higher in intubated patients due to application of uniform strategies on how to assess intubated patients [31].

Discussion

In this systematic review, 52 studies were identified that examined the reliability of the GCS in 13,142 patients. The studies varied with regard to the patient population, sample size, characteristics of the observers, study design and setting. The methodological quality was overall low. Good quality studies found the GCS to be adequately reliable when assessed by most key reliability measures (85 % of kappa values; 100 % of ICC values). Similar results were found in fair quality studies. However, despite this favorable overall conclusion, the estimates varied within and between studies, ranging from very

Table 4 Overview of reported factors that might influence reliability of the GCS

	Studies reporting the factor	Factor influences reliability	Influence not clear	No influence of factor
Observer related factors				
Education/training	[21, 22, 33, 35, 47, 58]	×	–	–
Level of experience	[21, 33–35, 47, 54, 56, 61, 64, 66, 76]	–	×	–
Type of profession[a]	[21, 22, 24, 26, 33, 40, 56, 57, 60–62, 65–67, 69–71, 76, 79, 81, 83]	–	–	×
Type of stimulus	[21, 51]	×	–	–
Patient related factors				
Consciousness level	[22, 24, 28, 32, 34, 48, 49, 52, 60, 61, 63, 64, 69, 71, 76, 77, 81, 82]	×	–	–
Type of pathology[b]	[25, 60, 64, 71, 76, 79]	–	×	–
Treatment[c]	[31, 62, 63, 70]	–	×	–

[a] The studies investigated following professions: (neuro)surgeon(s), resident(s), intensivist(s), neurologist(s), nurse(s), emergency medical technician(s), paramedic(s), student(s), physiotherapist, neuropsychologist(s)

[b] Type of pathologies studied included: traumatic brain injury, hemorrhagic stroke, ischemic stroke, subarachnoid hemorrhage, intoxication/poisoning, epilepsy, cardiovascular disease

[c] Sedation or intubation

poor to excellent reliability. The sum score was less reliable compared to the component scores, supporting existing reservations about the use of the sum score in the management of individual patients [6, 15, 37, 38]. This may reflect the fact that the sum score requires each of the three components to be assessed after which they are combined into one score, introducing four sources of potential observer variation. Moreover, the sum score has more possible scoring options (range 3–15), compared to the motor (range 1–6), verbal (range 1–5) and eye (range 1–4) components, implying a higher potential for disagreement. Similar, modest reliabilities for the sum score were found in the studies that focused on the GCS as part of the APACHE II [27, 28]. Although these studies concern ICU patients, other studies performed in the ICU performed in general much better. In particular, even higher overall kappa values were found in the selection of studies performed in the ICU, showing substantial agreement in 90.5 %, thereby justifying reliable use of the GCS in the ICU. We could not draw a clear conclusion regarding the intra-rater reliability due to the low number of studies, inconsistent use of reliability estimates, and low quality of studies.

Different reliability estimates were used across studies, with most having shortcomings. In particular, both percentage agreement and DR may overestimate true observer agreement [35, 39], and the DR is no longer considered as an appropriate reliability measure. The extent of disagreement is taken into account by the weighted kappa statistic, as the weighting results in a lower agreement when observers report larger differences [40]. Unfortunately, use of the weighted kappa was only seldom reported. ICC values have no absolute meaning, as they are strongly influenced by the heterogeneity of the population. Moreover, if the GCS is considered as an ordinal categorical variable, use of the ICC can be challenged. However, it may be argued that the sum score represents a continuous variable as its relationship with outcome is approximately linear [6]. Therefore, interpretation and combining the findings of primary studies is hampered and the precision with which a meaningful single estimate can be identified for the reliability is limited. However, this heterogeneity across studies does contribute to the generalizability of the findings of this systematic review.

To provide suggestions for optimizing reliability we analyzed factors that might influence results. We identified evidence that supports the effect of the following factors: training and education, type of stimulus and level of consciousness. Although the evidence did not support the influence of intubation and sedation on the reliability, GCS assessment in these treatment modalities is a commonly cited failing of the GCS in the ICU setting, as responses become untestable [12–14, 41]. Instructions on how to assess intubated patients can be expected to promote consistency [31]. Therefore, it is important to apply standardized approaches whenever a component is untestable. Teasdale et al. [6] recommend that a non-numerical designation 'NT' (not testable) should be assigned. The issue of untestable features is in particular relevant to the use of the GCS in aggregated ICU severity scores such as the APACHE II [3], the sequential organ failure assessment (SOFA) [4] and the simplified acute physiology score (SAPS) [5]. Pseudoscoring by averaging the testable scores or assuming a normal GCS score will affect the performance of these scoring systems. Various other options have been suggested to deal with untestable features, including use of the most reliable GCS score prior to sedation/intubation [42], imputing a score of one, and use of a linear regression model based upon scores of the other components [43]. Alternatively, the weighting of features included in aggregate score could be redefined to include the category 'untestable' as a separate category. We consider it a priority to develop

consensus on how best to deal with untestable components when entering the sum score into aggregated scoring systems.

Quality of studies

Across studies, the methodological quality ranged from poor to good, reflecting inadequate reporting and methodological flaws. This limits the strength of conclusions we can draw. The overall higher quality of studies conducted in more recent years reflects experience and the impact of guidelines for quality standards. Application of these standards to preceding studies led to a fairly high rate of ratings of poor/fair quality. This should perhaps not be considered as a proof of low quality, but a consequence of appropriate standards not being available at that time. We based our conclusions on higher quality studies and checked for reflection of the results in lower quality studies.

Relationship to previous work

The findings of this study extend those in previous reviews, in which a variety of findings have been reported. Koch and Linn [37] stated in their comprehensive review that the GCS scale is reliable and consistent for evaluating responsiveness and for predicting outcome of coma. In contrast, Baker et al. [11] concluded that it remains unclear whether the GCS has sufficient inter-rater reliability and emphasize that the evidence base is derived from inconsistent research methodologies, leading to a picture of ambiguity. Prasad [16] noted that reliability is good if no untestable features are present and observers are experienced. A more recent editorial stated that the GCS has repeatedly demonstrated surprisingly low inter-observer reliability. This opinion was, however, based on a review of only eight reliability studies and two review articles [12]. Also, Zuercher et al. [15] recognized that there is considerable inaccuracy in GCS scoring in daily practice as well as in clinical research and emphasize the need for consistent use of the GCS and quality improvement initiatives to increase the accuracy of scoring [44].

No previous study has tried to establish an overall estimate of reliability. Although this systematic review recognizes several conflicting findings among primary studies, it also shows that 81 % of all reported kappa values showed substantial agreement, which can be considered a proof of adequate clinical reliability. Consequently, this systematic review does not endorse the criticisms on the reliability of the GCS [12, 44]. Debate is ongoing about what level of reliability is acceptable for clinical care and health research. The classification of Landis and Koch is often applied, but may be too liberal because it refers to kappa scores as low as 0.41 as acceptable [39]. In this systematic review, we focused on levels of 0.6 and 0.7. The former is referred to by Landis and Koch as 'substantial' [18] and the latter by Terwee et al. as minimum standard for reliability [19].

Strengths and limitations

The strengths of this systematic review are that we employed a comprehensive search strategy, and followed accepted best practice [17] for key review tasks. However, it is possible that we missed unpublished data, because we did not search the grey literature. In addition, a greater depth of information could have been obtained by contacting the authors of primary studies to derive or clarify missing data.

Implications and recommendations

This study has implications for further reliability research and for clinical practice.

From the former perspective, the methodological flaws and inadequate reporting, reflected in the low quality of many studies, should be improved. In future research, observers and patients should be clearly characterized and sufficient numbers studied. A compatible approach to analysis should be used across studies, as outlined in guidelines developed for reporting reliability [45]. The Kappa statistic, while not without limitations [46], is currently the most widely applied reliability estimate for nominal measures and accompanying confidence intervals should be reported to facilitate meta-analysis.

In clinical practice, the overall reliability of the GCS seems to be adequate. However, "adequate" should not be considered sufficient. Standards for an important clinical monitoring instrument should be high. The broad range of reliability estimates reported in the literature indicates room for improvement. Endeavors to improve the reliability should be guided by an understanding of the factors that influence reliability. Awareness should be raised that reliability of the sum score is less than that of the components of the GCS, and this should be taken into consideration when using the sum score in disease severity scores or prediction models.

Conclusion

This systematic review identified a general lack of high quality studies and revealed considerable heterogeneity between studies. Despite these caveats, good quality

studies show adequate reliability of the GCS. The higher reliability in assessing the three components endorses their use over the sum score in describing individual patients. The findings of this study underscore the importance of efforts to improve reliability research in this field and emphasize the importance of continuing efforts to improve the reliability of the GCS in order to optimize its use in clinical practice. To this purpose, we present the following recommendations:

1. Ensure teaching and training in GCS to all new/inexperienced users across relevant disciplines.
2. Provide regular education and re-assurance of competence for experienced users.
3. Apply standardized stimuli to assess unresponsive patients.
4. Apply uniform strategies to deal with untestable features.
5. Report and communicate each of the three components of the GCS, rather than using the sum score.
6. Develop consensus on how to enter the sum score in aggregated ICU scoring systems.

Acknowledgments The authors would like to thank Sir Graham Teasdale, Emeritus Professor of Neurosurgery, University of Glasgow, UK, for the valuable discussions and very helpful contribution throughout the course of this work. This work was in part supported by the Framework 7 program of the European Union in the context of CENTER-TBI (Grant Number 602150-2).

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Teasdale G, Jennett B (1974) Assessment of coma and impaired consciousness. A practical scale. Lancet 2:81–84
2. Koziol J, Hacke W (1990) Multivariate data reduction by principal components with application to neurological scoring instruments. J Neurol 237:461–464
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13:818–829
4. Vincent JL, Moreno R, Takala J et al (1996) The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 22:707–710
5. Le Gall JR, Lemeshow S, Saulnier F (1993) A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 270:2957–2963
6. Teasdale G, Maas A, Lecky F et al (2014) The Glasgow Coma Scale at 40 years: standing the test of time. Lancet Neurol 13:844–854. doi: 10.1016/S1474-4422(14)70120-6
7. Murray GD, Butcher I, McHugh GS et al (2007) Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. J Neurotrauma 24:329–337
8. Mokkink LB, Terwee CB, Patrick DL et al (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 63:737–745. doi: 10.1016/j.jclinepi.2010.02.006
9. Fava GA, Tomba E, Sonino N (2012) Clinimetrics: the science of clinical measurements. Int J Clin Pract 66:11–15. doi: 10.1111/j.1742-1241.2011.02825.x
10. Sun S (2011) Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Methodol 11:145–163
11. Baker M (2008) Reviewing the application of the Glasgow Coma Scale: Does it have interrater reliability? Br J Neurosci Nurs 4:342–347. doi: 10.12968/bjnn.2008.4.7.30674
12. Green SM (2011) Cheerio, laddie! Bidding farewell to the Glasgow Coma Scale. Ann Emerg Med 58:427–430
13. Lowry M (1998) Emergency nursing and the Glasgow Coma Scale. Accid Emerg Nurs 6:143–148
14. Wijdicks EFM (2006) Clinical scales for comatose patients: the Glasgow Coma Scale in historical context and the new FOUR Score. Rev Neurol Dis 3:109–117
15. Zuercher M, Ummenhofer W, Baltussen A, Walder B (2009) The use of Glasgow Coma Scale in injury assessment: a critical review. Brain Inj 23:371–384. doi: 10.1080/02699050902926267
16. Prasad K (1996) The Glasgow Coma Scale: a critical appraisal of its clinimetric properties. J Clin Epidemiol 49:755–763
17. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 6:e1000097. doi: 10.1371/journal.pmed.1000097
18. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
19. Terwee CB, Bot SDM, de Boer MR et al (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34–42. doi: 10.1016/j.jclinepi.2006.03.012
20. Fleiss J (1986) The design and analysis of clinical experiments. Wiley, New York
21. Teasdale G, Knill-Jones R, van der Sande J (1978) Observer variability in assessing impaired consciousness and coma. J Neurol Neurosurg Psychiatry 41:603–610
22. Heron R, Davie A, Gillies R, Courtney M (2001) Interrater reliability of the Glasgow Coma Scale scoring among nurses in sub-specialties of critical care. Aust Crit Care 14:100–105
23. Nunnally JC, Bernstein IH (1994) Psychometric theory, 3rd edn. McGraw-Hill, New York

24. Menegazzi JJ, Davis EA, Sucov AN, Paris PM (1993) Reliability of the Glasgow Coma Scale when used by emergency physicians and paramedics. J Trauma 34:46–48
25. Hollander JE, Go S, Lowery DW et al (2003) Inter-rater reliability of criteria used in assessing blunt head injury patients for intracranial injuries. Acad Emerg Med 10:830–835
26. Kho ME, McDonald E, Stratford PW, Cook DJ (2007) Interrater reliability of APACHE II scores for medical-surgical intensive care patients: a prospective blinded study. Am J Crit Care 16:378–383
27. Wenner JB, Norena M, Khan N et al (2009) Reliability of intensive care unit admitting and comorbid diagnoses, race, elements of Acute Physiology and Chronic Health Evaluation II Score, and predicted probability of mortality in an electronic intensive care unit database. J Crit Care 24:401–407
28. Chen LM, Martin CM, Morrison TL, Sibbald WJ (1999) Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. Crit Care Med 27:1999–2004
29. Diringer MN, Edwards DF (1997) Does modification of the Innsbruck and the Glasgow Coma Scales improve their ability to predict functional outcome? Arch Neurol 54:606–611
30. Sadaka F, Patel D, Lakshmanan R (2012) The FOUR score predicts outcome in patients after traumatic brain injury. Neurocrit Care 16:95–101
31. Gill MR, Reiley DG, Green SM (2004) Inter-rater reliability of Glasgow Coma Scale scores in the emergency department. Ann Emerg Med 43:215–223
32. Feldman A, Hart KW, Lindsell CJ, McMullan JT (2015) Randomized controlled trial of a scoring aid to improve Glasgow Coma Scale scoring by emergency medical services providers. Ann Emerg Med 65(325–329):e2. doi: 10.1016/j.annemergmed.2014.07.454
33. Juarez VJ, Lyons M (1995) Interrater reliability of the Glasgow Coma Scale. J Neurosci Nurs 27:283–286
34. Rowley G, Fielding K (1991) Reliability and accuracy of the Glasgow Coma Scale with experienced and inexperienced users. Lancet 337:535–538
35. Fielding K, Rowley G (1990) Reliability of assessments by skilled observers using the Glasgow Coma Scale. Aust J Adv Nurs 7:13–17
36. Gill M, Martens K, Lynch EL et al (2007) Inter-rater reliability of 3 simplified neurologic scales applied to adults presenting to the emergency department with altered levels of consciousness. Ann Emerg Med 49:403–407
37. Koch D, Linn S (2000) The Glasgow Coma Scale and the challenge of clinimetrics. Int Med J 7:51–60
38. Sternbach GL (2000) The Glasgow coma scale. J Emerg Med 19:67–71
39. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Medica 22:276–282
40. Heard K, Bebarta VS (2004) Reliability of the Glasgow Coma Scale for the emergency department evaluation of poisoned patients. Hum Exp Toxicol 23:197–200
41. Le Roux P, Menon DK, Citerio G et al (2014) Consensus summary statement of the international multidisciplinary consensus conference on multimodality monitoring in neurocritical care : a statement for healthcare professionals from the Neurocritical Care Society and the European Society of Intensive Care Medicine. Intensive Care Med 40(9):1189–1209. doi: 10.1007/s00134-014-3369-6
42. Livingston BM, Mackenzie SJ, MacKirdy FN, Howie JC (2000) Should the pre-sedation Glasgow Coma Scale value be used when calculating acute physiology and chronic health evaluation scores for sedated patients? Scottish Intensive Care Society Audit Group. Crit Care Med 28:389–394
43. Rutledge R, Lentz CW, Fakhry S, Hunt J (1996) Appropriate use of the Glasgow Coma Scale in intubated patients: a linear regression prediction of the Glasgow Verbal Score from the Glasgow Eye and Motor scores. J Trauma 41:514–522
44. Zuercher M, Ummenhofer W, Baltussen A, Walder B (2009) The use of Glasgow Coma Scale in injury assessment: a critical review. Brain Inj 23:371–384. doi: 10.1080/02699050902926267
45. Kottner J, Audige L, Brorson S et al (2011) Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 64:96–106. doi: 10.1016/j.jclinepi.2010.03.002
46. Zhao X (2011) When to use Cohen's k, if Ever?. International Communication Association 2011 Conference, Boston, pp 1–30
47. Braakman R, Avezaat CJ, Maas AI et al (1977) Inter observer agreement in the assessment of the motor response of the Glasgow coma scale. Clin Neurol Neurosurg 80:100–106
48. Rimel RW, Jane JA, Edlich RF (1979) An injury severity scale for comprehensive management of central nervous system trauma. JACEP 8:64–67
49. Lindsay KW, Teasdale GM, Knill-Jones RP (1983) Observer variability in assessing the clinical features of subarachnoid hemorrhage. J Neurosurg 58:57–62
50. Stanczak DE, White JG 3rd, Gouview WD et al (1984) Assessment of level of consciousness following severe neurological insult. A comparison of the psychometric qualities of the Glasgow Coma Scale and the comprehensive level of Consciousness Scale. J Neurosurg 60:955–960. doi: 10.3171/jns.1984.60.5.0955
51. Starmark JE, Heath A (1988) Severity grading in self-poisoning. Hum Toxicol 7:551–555
52. Starmark JE, Stalhammar D, Holmgren E, Rosander B (1988) A comparison of the Glasgow Coma Scale and the Reaction Level Scale (RLS85). J Neurosurg 69:699–706
53. Tesseris J, Pantazidis N, Routsi C, Fragoulakis D (1991) A comparative study of the Reaction Level Scale (RLS85) with Glasgow Coma Scale (GCS) and Edinburgh-2 Coma Scale (modified) (E2CS(M)). Acta Neurochir Wien 110:65–76
54. Ellis A, Cavanagh SJ (1992) Aspects of neurosurgical assessment using the Glasgow Coma Scale. Intensive Crit Care Nurs 8:94–99
55. Oshiro EM, Walter KA, Piantadosi S et al (1997) A new subarachnoid hemorrhage grading system based on the Glasgow Coma Scale: a comparison with the Hunt and Hess and World Federation of Neurological Surgeons Scales in a clinical series. Neurosurgery 41:140–147
56. Crossman J, Bankes M, Bhan A, Crockard HA (1998) The Glasgow Coma Score: reliable evidence? Injury 29:435–437
57. Wijdicks EF, Kokmen E, O'Brien PC (1998) Measurement of impaired consciousness in the neurological intensive care unit: a new test. J Neurol Neurosurg Psychiatry 64:117–119
58. Lane PL, Baez AA, Brabson T et al (2002) Effectiveness of a Glasgow Coma Scale instructional video for EMS providers. Prehosp Disaster Med 17:142–146
59. Ely EW, Truman B, Shintani A et al (2003) Monitoring sedation status over time in ICU patients: reliability and validity of the Richmond Agitation-Sedation Scale (RASS). JAMA 289:2983–2991

60. Wijdicks EF, Bamlet WR, Maramattom BV et al (2005) Validation of a new Coma Scale: the FOUR score. Ann Neurol 58:585–593
61. Holdgate A, Ching N, Angonese L (2006) Variability in agreement between physicians and nurses when measuring the Glasgow Coma Scale in the emergency department limits its clinical usefulness. Emerg Med Australas 18:379–384
62. Baez AA, Giraldez EM, De Pena JM (2007) Precision and reliability of the Glasgow Coma Scale score among a cohort of Latin American prehospital emergency care providers. Prehosp Disaster Med 22:230–232
63. Kerby JD, MacLennan PA, Burton JN et al (2007) Agreement between prehospital and emergency department Glasgow Coma Scores. J Trauma 63:1026–1031
64. Wolf CA, Wijdicks EF, Bamlet WR, McClelland RL (2007) Further validation of the FOUR Score Coma Scale by intensive care nurses. Mayo Clin Proc 82:435–438
65. Nassar AP Jr, Neto RCP, de Figueiredo WB, Park M (2008) Validity, reliability and applicability of Portuguese versions of Sedation-Agitation Scales among critically ill patients. Sao Paulo Med J 126:215–219
66. Akavipat P (2009) Endorsement of the FOUR Score for consciousness assessment in neurosurgical patients. Neurol Med Chir Tokyo 49:565–571
67. Iyer VN, Mandrekar JN, Danielson RD et al (2009) Validity of the FOUR Score Coma Scale in the medical intensive care unit. Mayo Clin Proc 84:694–701
68. Ryu WHA, Feinstein A, Colantonio A et al (2009) Early identification and incidence of mild TBI in Ontario. Can J Neurol Sci 36:429–435
69. Stead LG, Wijdicks EF, Bhagra A et al (2009) Validation of a new Coma Scale, the FOUR Score, in the emergency department. Neurocrit Care 10:50–54
70. Fischer M, Ruegg S, Czaplinski A et al (2010) Inter-rater reliability of the Full Outline of Unresponsiveness Score and the Glasgow Coma Scale in critically ill patients: a prospective observational study. Crit Care 14:R64
71. Idrovo L, Fuentes B, Medina J et al (2010) Validation of the FOUR Score (Spanish Version) in acute stroke: an interobserver variability study. Eur Neurol 63:364–369
72. Necioglu Orken D, Kocaman Sagduyu A, Sirin H et al (2010) Reliability of the Turkish version of a new Coma Scale: FOUR Score. Balk Med J 27:28–31
73. Ashkenazy S, DeKeyser-Ganz F (2011) Assessment of the reliability and validity of the Comfort Scale for adult intensive care patients. Heart Lung 40:e44–e51
74. Bruno MA, Ledoux D, Lambermont B et al (2011) Comparison of the FOUR and Glasgow Liege Scale/Glasgow Coma Scale in an intensive care unit population. Neurocrit Care 15:447–453
75. Gujjar AR, Nandagopal R, Jacob PC et al (2011) FOUR Score–a new Coma Score: inter-observer reliability and relation to outcome in critically ill medical patients. Eur J Neurol 18:441
76. Kevric J, Jelinek GA, Knott J, Weiland TJ (2011) Validation of the Full Outline of Unresponsiveness (FOUR) Scale for conscious state in the emergency department: comparison against the Glasgow Coma Scale. Emerg Med J 28:486–490
77. Namiki J, Yamazaki M, Funabiki T, Hori S (2011) Inaccuracy and misjudged factors of Glasgow Coma Scale scores when assessed by inexperienced physicians. Clin Neurol Neurosurg 113:393–398
78. Patel D, Sadaka F, Lakshmanan R (2011) The FOUR score predicts outcome in patients after traumatic brain injury. Neurocrit Care 15:S238
79. Takahashi C, Okudera H, Origasa H et al (2011) A simple and useful coma scale for patients with neurologic emergencies: the Emergency Coma Scale. Am J Emerg Med 29:196–202
80. Winship C, Williams B, Boyle M (2011) Assessment of the Glasgow Coma Scale: a pilot study examining the accuracy of paramedic undergraduates. Australas J Paramed 10:11
81. Marcati E, Ricci S, Casalena A et al (2012) Validation of the Italian version of a new Coma Scale: the Four Score. Intern Emerg Med 7:145–152
82. Winship C, Williams B, Boyle MJ (2012) Should an alternative to the Glasgow Coma Scale be taught to paramedic students? Emerg Med J 30:e19
83. Benítez-Rosario MA, Castillo-Padrós M, Garrido-Bernet B et al (2013) Appropriateness and reliability testing of the modified richmond Agitation-Sedation Scale in Spanish patients with advanced cancer. J Pain Symptom Manag 45:1112–1119
84. Dinh MM, Oliver M, Bein K et al (2013) Level of agreement between prehospital and emergency department vital signs in trauma patients. Emerg Med Australas 25:457–463
85. Gujjar AR, Jacob PC, Nandhagopal R et al (2013) FOUR score and Glasgow Coma Scale in medical patients with altered sensorium: interrater reliability and relation to outcome. J Crit Care 28(316):e1–e8